Estimation of a signal matrix for high-dimensional non-Gaussian data (Statistical Inference on Divergence Measures and Its Related Topics)

AUTHOR(S):

矢田, 和善; 青嶋, 誠

数理解析研究所講究録
第 1999 巻 2016 年 36-46

Estimation of a signal matrix for high-dimensional non-Gaussian data

筑波大学・数理物質系　矢田　和善
Kazuyoshi Yata
Institute of Mathematics
University of Tsukuba

筑波大学・数理物質系　青嶋　誠
Makoto Aoshima
Institute of Mathematics
University of Tsukuba

Abstract: We consider the problem of recovering a low-rank signal matrix in high-dimensional situations. We first consider the conventional PCA to recover the signal matrix and show that the estimation of the signal matrix holds consistency properties under severe conditions. The conventional PCA is heavily subjected to the noise. In order to reduce the noise, we consider using the noise reduction (NR) methodology to recover the signal matrix and show that the estimation by the NR method improves the error rate of the conventional PCA effectively. We also apply the cross-data-matrix (CDM) methodology to recover the signal matrix and propose a new estimation of the signal matrix. We show that the proposed estimation by the CDM method performs well for high-dimensional non-Gaussian data.

Key words and phrases: Cross-data-matrix methodology, HDLSS, Large p small n, Noise-reduction methodology.

1 Introduction

In this paper, we address the problem of recovering an unknown $d \times n$ low-rank matrix, $\boldsymbol{A} = [\boldsymbol{a}_1, ..., \boldsymbol{a}_n]$. $\boldsymbol{A}$ is called the signal matrix. Let $r = \mathrm{rank}(\boldsymbol{A})$. We assume r $(< \min\{d, n\})$ is fixed. Suppose we have a $d \times n$ data matrix, $\boldsymbol{X} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_n]$, where

$$\boldsymbol{X} = \sqrt{n}\boldsymbol{A} + \boldsymbol{W}. \tag{1}$$

Here, $\boldsymbol{W} = [\boldsymbol{w}_1, ..., \boldsymbol{w}_n]$ is a $d \times n$ noise matrix, where $\boldsymbol{w}_j$, $j = 1, ..., n$, are independent and identically distributed (i.i.d.) as a d-dimensional distribution with mean zero and covariance matrix $\boldsymbol{\Sigma}_W$ $(\neq \boldsymbol{O})$. Note that $\boldsymbol{x}_j - \sqrt{n}\boldsymbol{a}_j$, $j = 1, ..., n$,

are i.i.d. Let $\Sigma_A = \boldsymbol{A}\boldsymbol{A}^T$. Then, it holds that $E(\boldsymbol{X}\boldsymbol{X}^T)/n = \Sigma_A + \Sigma_W$ $(= \Sigma$, say). Shabalin and Nobel [6] considered (1) in a high-dimensional setting, where the data dimension d and the sample size n increase at the same rate, i.e. $n/d \to c > 0$. They assumed that the elements of $\boldsymbol{W}$ are i.i.d. normal random variables. We note that the conditions such as "$n/d \to c > 0$" and the Gaussianity of the noise are often strict in real high-dimensional analyses. Yata and Aoshima [10] considered (1) in high-dimension, low-sample-size (HDLSS) settings without assuming those conditions.

The eigen-decomposition of Σ_W is given by $\Sigma_W = \boldsymbol{U}_W \boldsymbol{\Lambda}_W \boldsymbol{U}_W^T$, where $\boldsymbol{\Lambda}_W$ is a diagonal matrix of eigenvalues, $\lambda_{1(W)} \geq \cdots \geq \lambda_{d(W)} (\geq 0)$, and $\boldsymbol{U}_W$ is an orthogonal matrix of the corresponding eigenvectors. Let $\boldsymbol{W} = \boldsymbol{U}_W \boldsymbol{\Lambda}_W^{1/2} \boldsymbol{Z}$. Then, $\boldsymbol{Z}$ is a $d \times n$ sphered data matrix from a distribution with the identity covariance matrix. Here, we write $\boldsymbol{Z} = [\boldsymbol{z}_1, ..., \boldsymbol{z}_d]^T$ and $\boldsymbol{z}_j = (z_{j1}, ..., z_{jn})^T$, $j = 1, ..., d$. Note that $E(z_{jk} z_{j'k}) = 0$ $(j \neq j')$ and $\mathrm{Var}(\boldsymbol{z}_j) = \boldsymbol{I}_n$, where $\boldsymbol{I}_n$ is the n-dimensional identity matrix. We assume that the fourth moments of each variable in $\boldsymbol{Z}$ are uniformly bounded. The singular value decomposition of $\boldsymbol{A}$ is given by

$$\boldsymbol{A} = \sum_{j=1}^{r} \lambda_{j(A)}^{1/2} \boldsymbol{u}_{j(A)} \boldsymbol{v}_{j(A)}^T,$$

where $\lambda_{1(A)}^{1/2} \geq \cdots \geq \lambda_{r(A)}^{1/2}$ (> 0) are singular values of $\boldsymbol{A}$ and $\boldsymbol{u}_{j(A)}$ (or $\boldsymbol{v}_{j(A)}$) denotes a unit left- (or right-) singular vector corresponding to $\lambda_{j(A)}^{1/2}$ $(j = 1, ..., r)$. Note that $\Sigma_A = \sum_{j=1}^{r} \lambda_{j(A)} \boldsymbol{u}_{j(A)} \boldsymbol{u}_{j(A)}^T$. Also, note that $\lambda_{j(A)}$s depend not only on d but also on n. In this paper, we assume the following model.

$$\lim_{d \to \infty} \frac{\mathrm{tr}(\Sigma_W^2)}{\lambda_{r(A)}^2} = 0 \quad \text{when } n \text{ is fixed or } n \to \infty. \tag{2}$$

The model (2) is a special case of the power spiked model given by Yata and Aoshima [9]. Murayama et al. [5] considered the estimation of $\boldsymbol{A}$ for a special case of (2). When $r \geq 2$, we assume that $\lambda_{j(A)}$s are distinct in the sense that

$$\liminf_{d \to \infty} \frac{\lambda_{j(A)}}{\lambda_{j'(A)}} > 1 \quad \text{when } n \text{ is fixed or } n \to \infty \text{ for all } j < j' \ (\leq r).$$

The sample covariance matrix is given by $\boldsymbol{S} = n^{-1} \boldsymbol{X}\boldsymbol{X}^T$. We consider the dual sample covariance matrix defined by $\boldsymbol{S}_D = n^{-1} \boldsymbol{X}^T \boldsymbol{X}$. Let $m = \min\{d, n\}$. Note that $\boldsymbol{S}_D$ and $\boldsymbol{S}$ share non-zero eigenvalues and $\mathrm{rank}(\boldsymbol{S}) = \mathrm{rank}(\boldsymbol{S}_D) \leq m$. Let $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_m \geq 0$ be the eigenvalues of $\boldsymbol{S}_D$. The eigen-decompositions of $\boldsymbol{S}$ and $\boldsymbol{S}_D$ are given by $\boldsymbol{S} = \sum_{j=1}^{m} \hat{\lambda}_j \hat{\boldsymbol{u}}_j \hat{\boldsymbol{u}}_j^T$ and $\boldsymbol{S}_D = \sum_{j=1}^{m} \hat{\lambda}_j \hat{\boldsymbol{v}}_j \hat{\boldsymbol{v}}_j^T$, where $\hat{\boldsymbol{u}}_j$ (or $\hat{\boldsymbol{v}}_j$) denotes a unit left- (or right-) singular vector of $\boldsymbol{X}/n^{1/2}$ corresponding to $\hat{\lambda}_j^{1/2}$. Note that $\hat{\boldsymbol{u}}_j$ can be calculated by $\hat{\boldsymbol{u}}_j = (n\hat{\lambda}_j)^{-1/2} \boldsymbol{X} \hat{\boldsymbol{v}}_j$ from the fact that $\boldsymbol{X}/n^{1/2} = \sum_{j=1}^{m} \hat{\lambda}_j^{1/2} \hat{\boldsymbol{u}}_j \hat{\boldsymbol{v}}_j^T$.

Jung and Marron [4] investigated the inconsistency of the eigenvalues and eigenvectors of the sample covariance matrix for HDLSS data. Yata and Aoshima [8] gave consistent estimators for both the eigenvalues and eigenvectors together with the principal component (PC) scores by developing the *noise-reduction (NR) methodology.* Moreover, Yata and Aoshima [7] created a new principal component analysis (PCA) called the *cross-data-matrix (CDM) methodology* that is applicable to constructing an unbiased estimator in nonparametric settings. Aoshima and Yata [1] developed a variety of high-dimensional statistical inference by using the cross-data-matrix methodology. See Aoshima and Yata [2, 3] for a review covering this field of research.

In this paper, we consider the problem of recovering the signal matrix $\boldsymbol{A}$ in high-dimensional settings. In Sections 2 and 3, we introduce the results of Yata and Aoshima [10]. In Section 2, we consider using the conventional PCA to recover $\boldsymbol{A}$ and show that the estimation of $\boldsymbol{A}$ holds consistency properties under severe conditions. In Section 3, we consider the noise reduction (NR) methodology by Yata and Aoshima [8] in (1) and apply it to recovering $\boldsymbol{A}$. We show that the estimation of $\boldsymbol{A}$ by the NR method holds the consistency properties under mild conditions and improves the error rate of the conventional PCA. In Section 4, we consider the cross-data-matrix (CDM) methodology by Yata and Aoshima [7] in (1) and apply it to recovering $\boldsymbol{A}$. We show that the estimation of $\boldsymbol{A}$ by the CDM method performs well for high-dimensional non-Gaussian data.

2 Estimation of the signal matrix by conventional PCA

In this section, we consider recovering the signal matrix $\boldsymbol{A}$ by using the conventional PCA in high-dimensional settings such as $d \to \infty$ either when n is fixed or $n \to \infty$. We reconstruct $\boldsymbol{A}$ by using $\hat{\lambda}_j$s, $\hat{\boldsymbol{u}}_j$s and $\hat{\boldsymbol{v}}_j$s. We assume $\hat{\boldsymbol{u}}_j^T \boldsymbol{u}_{j(A)} \geq 0$ and $\hat{\boldsymbol{v}}_j^T \boldsymbol{v}_{j(A)} \geq 0$ for all j $(\leq r)$ without loss of generality.

We consider the following conditions when $d \to \infty$ while n is fixed or $n \to \infty$:

$$\textbf{(C-i)} \quad \frac{\sum_{s,t=1}^{d} \lambda_{s(W)} \lambda_{t(W)} E\{(z_{sk}^2 - 1)(z_{tk}^2 - 1)\}}{n \lambda_{r(A)}^2} = o(1);$$

$$\textbf{(C-ii)} \quad \frac{\mathrm{tr}(\boldsymbol{\Sigma}_W)}{n \lambda_{r(A)}} = o(1).$$

Remark 1. *We note that* $z_{1k}, ..., z_{dk}$ $(k = 1, ..., n)$ *are independent when* $\boldsymbol{W}$ *is Gaussian. Then, it holds that*

$$\sum_{s,t=1}^{d} \lambda_{s(W)} \lambda_{t(W)} E\{(z_{sk}^2 - 1)(z_{tk}^2 - 1)\} = O\{tr(\boldsymbol{\Sigma}_W^2)\}, \tag{3}$$

so that (C-i) holds under (2) when $\boldsymbol{W}$ *is Gaussian or* $z_{1k}, ..., z_{dk}$ $(k = 1, ..., n)$ *are independent.*

If (3) holds, (C-i) is met even when n is fixed. Let $\kappa_j = \mathrm{tr}(\boldsymbol{\Sigma}_W)/(n\lambda_{j(A)})$ for $j = 1, ..., r$. Yata and Aoshima [10] gave the following results.

Theorem 2.1 ([10]). *Under (C-i), it holds that for* $j = 1, ..., r$

$$\frac{\hat{\lambda}_j}{\lambda_{j(A)}} = 1 + \kappa_j + o_p(1), \quad \hat{\boldsymbol{u}}_j^T \boldsymbol{u}_{j(A)} = (1 + \kappa_j)^{-1/2} + o_p(1)$$
$$\text{and} \quad \hat{\boldsymbol{v}}_j^T \boldsymbol{v}_{j(A)} = 1 + o_p(1)$$

as $d \to \infty$ *either when* n *is fixed or* $n \to \infty$.

Corollary 2.1 ([10]). *Under (C-i) and (C-ii), it holds that for* $j = 1, ..., r$

$$\frac{\hat{\lambda}_j}{\lambda_{j(A)}} = 1 + o_p(1) \quad \text{and} \quad \hat{\boldsymbol{u}}_j^T \boldsymbol{u}_{j(A)} = 1 + o_p(1)$$

as $d \to \infty$ *either when* n *is fixed or* $n \to \infty$.

Note that $\hat{\boldsymbol{v}}_j$s hold the consistency property without (C-ii) contrary to $\hat{\lambda}_j$s and $\hat{\boldsymbol{u}}_j$s. Based on the theoretical background, we consider recovering the signal matrix $\boldsymbol{A}$ by $\hat{\boldsymbol{A}} = \sum_{i=1}^{r} \hat{\lambda}_i^{1/2} \hat{\boldsymbol{u}}_i \hat{\boldsymbol{v}}_i^T$. Yata and Aoshima [10] discussed the choice of r in $\hat{\boldsymbol{A}}$. We define a loss function by

$$L(\hat{\boldsymbol{A}}|\boldsymbol{A}) = ||\hat{\boldsymbol{A}} - \boldsymbol{A}||_F^2,$$

where $||\cdot||_F$ denotes the Frobenius norm. Let $\psi = \mathrm{tr}(\boldsymbol{\Sigma}_W)/n$. Then, Yata and Aoshima [10] gave the following results.

Theorem 2.2 ([10]). *Under (C-i), it holds that*

$$L(\hat{\boldsymbol{A}}|\boldsymbol{A}) = r\psi + o_p(\lambda_{r(A)})$$

as $d \to \infty$ *either when* n *is fixed or* $n \to \infty$.

Remark 2. *If (3) holds, Theorem 2.2 is claimed even when* n *is fixed.*

Corollary 2.2 ([10]). *Under (C-i) and (C-ii), it holds that*

$$L(\hat{\boldsymbol{A}}|\boldsymbol{A}) = o_p(\lambda_{r(A)})$$

as $d \to \infty$ *either when* n *is fixed or* $n \to \infty$.

From Theorem 2.2, if (C-ii) does not hold, the loss of $\hat{\boldsymbol{A}}$ becomes $r\mathrm{tr}(\boldsymbol{\Sigma}_W)/n$ asymptotically. In order to reduce the noise, we consider using the NR method to recover the signal matrix in Section 3.

3 Estimation of the signal matrix by NR method

In this section, we consider applying the *noise-reduction (NR) methodology* by Yata and Aoshima [8] to recover the signal matrix $\boldsymbol{A}$. By using the NR method, we obtain an estimator of $\lambda_{j(A)}$ as

$$\acute{\lambda}_{j(r)} = \hat{\lambda}_j - \frac{\mathrm{tr}(\boldsymbol{S}_D) - \sum_{i=1}^{r} \hat{\lambda}_i}{n-r} \quad (j = 1, ..., r). \tag{4}$$

Note that the second term in (4) is an estimator of ψ. Then, Yata and Aoshima [10] gave the following result.

Theorem 3.1 ([10]). *Under (C-i), it holds that for* $j = 1, ..., r$

$$\frac{\acute{\lambda}_{j(r)}}{\lambda_{j(A)}} = 1 + o_p(1)$$

as $d \to \infty$ *either when* n *is fixed or* $n \to \infty$.

From Theorem 3.1, $\acute{\lambda}_{j(r)}$ holds the consistency property without (C-ii). Remember that $\hat{\lambda}_j$ requires (C-ii) to hold the consistency property.

We consider recovering $\boldsymbol{A}$ by $\acute{\boldsymbol{A}} = \sum_{i=1}^{r} \acute{\lambda}_{i(r)}^{1/2} \hat{\boldsymbol{u}}_i \hat{\boldsymbol{v}}_i^T$. Yata and Aoshima [10] discussed the choice of r in $\acute{\boldsymbol{A}}$. Let

$$\delta_i = \boldsymbol{u}_{i(A)}^T \boldsymbol{W} \boldsymbol{v}_{i(A)} / (n\lambda_{i(A)})^{1/2} \quad \text{for } i = 1, ..., r.$$

For the loss function by $L(\acute{\boldsymbol{A}}|\boldsymbol{A}) = ||\acute{\boldsymbol{A}} - \boldsymbol{A}||_F^2$, Yata and Aoshima [10] gave the following results.

Theorem 3.2. *Under (C-i), it holds that*

$$L(\acute{\boldsymbol{A}}|\boldsymbol{A}) = 2\sum_{i=1}^{r} \lambda_{i(A)}(1+\delta_i)\Big(1 - \frac{1+\delta_i}{(1+\kappa_i+2\delta_i)^{1/2}}\Big) + o_p(\lambda_{r(A)})$$

$$\text{and } \delta_i = o_p\{(\lambda_{r(A)}/\lambda_{i(A)})^{1/2}\} \quad \text{for } i = 1, ..., r$$

as $d \to \infty$ *either when* n *is fixed or* $n \to \infty$.

Remark 3. *If (3) holds, Theorem 3.2 is claimed even when* n *is fixed.*

Corollary 3.1 ([10]). *Under (C-i) and (C-ii), it holds that*

$$L(\acute{\boldsymbol{A}}|\boldsymbol{A}) = o_p(\lambda_{r(A)})$$

as $d \to \infty$ *either when* n *is fixed or* $n \to \infty$.

From Theorems 2.2 and 3.2, we compare $2\lambda_{i(A)}\{1 - 1/(1+\kappa_i)^{1/2}\}$ with ψ $(= \lambda_{i(A)}\kappa_i)$ by noting $\delta_i = o_p(1)$. It holds that $2\{1 - 1/(1+\kappa_i)^{1/2}\} < \kappa_i$ $(i = 1, ..., r)$ for any $\kappa_i > 0$, so that $L(\acute{\boldsymbol{A}}|\boldsymbol{A})$ is smaller than $L(\hat{\boldsymbol{A}}|\boldsymbol{A})$ asymptotically. Thus, $\acute{\boldsymbol{A}}$ improves the loss of $\hat{\boldsymbol{A}}$.

4 Estimation of the signal matrix by CDM method

In this section, we consider applying the *cross-data-matrix (CDM) methodology* by Yata and Aoshima [7] to recover the signal matrix $\boldsymbol{A}$.

Let $n_1 = \lceil n/2 \rceil$ and $n_2 = n - n_{(1)}$, where $\lceil x \rceil$ denotes the smallest integer $\geq x$. We assume $n_2 \geq r$. Let $\boldsymbol{X}_1 = [\boldsymbol{x}_1, ..., \boldsymbol{x}_{n_1}]$ and $\boldsymbol{X}_2 = [\boldsymbol{x}_{n_1+1}, ..., \boldsymbol{x}_n]$. We define

$$\boldsymbol{S}_{D(1)} = (n_1 n_2)^{-1/2} \boldsymbol{X}_1^T \boldsymbol{X}_2.$$

Let $m_0 = \min\{d, n_2\}$. The singular value decomposition of $\boldsymbol{S}_{D(1)}$ is given by

$$\boldsymbol{S}_{D(1)} = \sum_{j=1}^{m_0} \tilde{\lambda}_j \tilde{\boldsymbol{v}}_{1j} \tilde{\boldsymbol{v}}_{2j}^T,$$

where $\tilde{\lambda}_j \geq \cdots \geq \tilde{\lambda}_{m_0} (\geq 0)$ denote singular values of $\boldsymbol{S}_{D(1)}$ and $\tilde{\boldsymbol{v}}_{1j}$ (or $\tilde{\boldsymbol{v}}_{2j}$) denotes a unit left- (or right-) singular vector corresponding to $\tilde{\lambda}_j$ $(j = 1, ..., m_0)$. Let $\boldsymbol{A}_1 = [\boldsymbol{a}_1, ..., \boldsymbol{a}_{n_1}]$, $\boldsymbol{A}_2 = [\boldsymbol{a}_{n_1+1}, ..., \boldsymbol{a}_n]$, $\boldsymbol{W}_1 = [\boldsymbol{w}_1, ..., \boldsymbol{w}_{n_1}]$ and $\boldsymbol{W}_2 = [\boldsymbol{w}_{n_1+1}, ..., \boldsymbol{w}_n]$. Then, we write that

$$\boldsymbol{X}_i = \sqrt{n} \boldsymbol{A}_i + \boldsymbol{W}_i, \quad i = 1, 2.$$

Let $\boldsymbol{v}_{j(A)} = (\boldsymbol{v}_{1j(A)}^T, \boldsymbol{v}_{2j(A)}^T)^T$ for $j = 1, ..., r$, where $\boldsymbol{v}_{ij(A)} \in \mathrm{R}^{n_i}$. Note that $||\boldsymbol{v}_{1j(A)}||^2 + ||\boldsymbol{v}_{2j(A)}||^2 = ||\boldsymbol{v}_{j(A)}||^2 = 1$ for $j = 1, ..., r$, where $|| \cdot ||$ denotes the Euclidean norm. Then, we write that

$$\boldsymbol{A}_i = \sum_{j=1}^{r} \lambda_{j(A)}^{1/2} \boldsymbol{u}_{j(A)} \boldsymbol{v}_{ij(A)}^T, \quad i = 1, 2.$$

Hereafter, we assume that

$$\limsup_{m \to \infty} \frac{\lambda_{1(A)}}{\lambda_{r(A)}} < \infty \quad \text{and} \quad ||\boldsymbol{v}_{ij(A)}||^2 = 1/2 + o(1) \quad \text{as } m \to \infty \text{ for all } i, j. \tag{5}$$

We assume $\tilde{\boldsymbol{v}}_{ij}^T \boldsymbol{v}_{ij(A)} \geq 0$ for all i, j without loss of generality. Then, we have the following result.

Theorem 4.1. *It holds that for* $j = 1, ..., r$

$$\frac{\tilde{\lambda}_j}{\lambda_{j(A)}} = 1 + o_p(1) \quad \textit{and} \quad \sqrt{2} \tilde{\boldsymbol{v}}_{ij}^T \boldsymbol{v}_{ij(A)} = 1 + o_p(1), \quad i = 1, 2,$$

as $m \to \infty$.

From Theorem 4.1, $\tilde{\lambda}_j$ holds the consistency property without (C-i) and (C-ii). Remember that $\acute{\lambda}_{j(r)}$ requires (C-i) to hold the consistency property.

Next, we consider estimation of $\boldsymbol{u}_{j(A)}$s by using the CDM methodology. Let $\tilde{\boldsymbol{u}}_{ij} = (n_i \tilde{\lambda}_j)^{-1/2} \boldsymbol{X}_i \tilde{\boldsymbol{v}}_{ij}$, $i = 1, 2$. Then, we have the following result.

Theorem 4.2. *It holds that for* $j = 1, ..., r$

$$\tilde{\boldsymbol{u}}_{ij}^T \boldsymbol{u}_{j(A)} = 1 + o_p(1), \quad i = 1, 2,$$

as $m \to \infty$.

Let $\tilde{\boldsymbol{u}}_{ij(*)} = \tilde{\boldsymbol{u}}_{ij}/||\tilde{\boldsymbol{u}}_{ij}||$ for all i, j. Let $\tilde{\boldsymbol{A}}_i = \sum_{j=1}^r \tilde{\lambda}_j^{1/2} \tilde{\boldsymbol{u}}_{ij(*)} \tilde{\boldsymbol{v}}_{ij}^T/\sqrt{2}$ for $i = 1, 2$. We consider recovering $\boldsymbol{A}$ by $\tilde{\boldsymbol{A}} = [\tilde{\boldsymbol{A}}_1, \tilde{\boldsymbol{A}}_2]$. From Theorems 4.1 and 4.2, we expect that $\tilde{\boldsymbol{A}}$ performs well for high-dimensional non-Gaussian data. In Section 5, we examine the performance of $\tilde{\boldsymbol{A}}$ with the help of numerical simulations.

5 Simulations

In this section, we give numerical comparisons of $\hat{\boldsymbol{A}}$, $\acute{\boldsymbol{A}}$ and $\tilde{\boldsymbol{A}}$.

We set $d = 2^t$, $t = 4, ..., 10$, $n = 15$, $r = 3$, $\boldsymbol{\Sigma}_A = \mathrm{diag}(\lambda_{1(A)}, \lambda_{2(A)}, \lambda_{3(A)}, 0, ..., 0)$ with $(\lambda_{1(A)}, \lambda_{2(A)}, \lambda_{3(A)}) = (d/5, d/15, d/45)$ and $\boldsymbol{\Sigma}_W = (0.3^{|i-j|^{1/3}})$. Note that $\mathrm{tr}(\boldsymbol{\Sigma}_W) = d$. We considered three cases:

(a) $\boldsymbol{w}_k$s are i.i.d. as a d-variate normal distribution, $N_d(\boldsymbol{0}, \boldsymbol{\Sigma}_W)$ with mean zero and covariance matrix $\boldsymbol{\Sigma}_W$;

(b) $\boldsymbol{w}_k$s are i.i.d. as a d-variate t-distribution, $t_d(\boldsymbol{0}, \boldsymbol{\Sigma}_W, \nu)$ with mean zero, covariance matrix $\boldsymbol{\Sigma}_W$ and degrees of freedom $\nu = 10$;

(c) $\boldsymbol{w}_k$s are i.i.d. as a d-variate t-distribution, $t_d(\boldsymbol{0}, \boldsymbol{\Sigma}_W, \nu)$ with $\nu = 30$.

Let $F(\boldsymbol{M}) = ||\boldsymbol{M} - \boldsymbol{A}||_F^2/d$ for any $d \times n$ matrix, $\boldsymbol{M}$. The findings were obtained by averaging the outcomes from 2000 ($= K$, say) replications. Under a fixed scenario, suppose that the k-th replication ends with estimates, $F(\hat{\boldsymbol{A}})_k$, $F(\acute{\boldsymbol{A}})_k$ and $F(\tilde{\boldsymbol{A}})_k$, for $k = 1, ..., K$. Let us simply write $\hat{F} = K^{-1} \sum_{k=1}^K F(\hat{\boldsymbol{A}}_r)_k$, $\acute{F} = K^{-1} \sum_{k=1}^K F(\acute{\boldsymbol{A}})_k$ and $\tilde{F} = K^{-1} \sum_{k=1}^K F(\tilde{\boldsymbol{A}})_k$. We also considered the Monte Carlo variability. Let $\mathrm{var}(\hat{F}) = (K-1)^{-1} \sum_{k=1}^K (F(\hat{\boldsymbol{A}})_k - \hat{F})^2$, $\mathrm{var}(\acute{F}) = (K-1)^{-1} \sum_{k=1}^K (F(\acute{\boldsymbol{A}})_k - \acute{F})^2$ and $\mathrm{var}(\tilde{F}) = (K-1)^{-1} \sum_{k=1}^K (F(\tilde{\boldsymbol{A}})_k - \tilde{F})^2$. Figure 1 shows the behaviors of $(\hat{F}, \acute{F}, \tilde{F})$ and $(\mathrm{var}(\hat{F}), \mathrm{var}(\acute{F}), \mathrm{var}(\tilde{F}))$ for (a), (b) and (c).

We observed that the NR method and the CDM method give more preferable performances compared to the conventional PCA. It seems that the NR method performs better than the CDM method for (a). Note that $t_d(\boldsymbol{0}, \boldsymbol{\Sigma}_W, \nu) \Rightarrow N_d(\boldsymbol{0}, \boldsymbol{\Sigma}_W)$ as $\nu \to \infty$. When $\nu = 10$, the NR method seems not to give a feasible estimation. This is probably because $\nu = 10$ is not large enough for $\boldsymbol{W}$ to satisfy (C-i). On the other hand, the CDM method does not require (C-i). As observed in Figure 2, the CDM method seems to perform well even when $\nu = 10$.

A Appendix

Throughout, let $\boldsymbol{e}_{in_i} = (e_{i1}, ..., e_{in_i})^T$, $i = 1, 2$, be arbitrary unit random vectors. Let $\boldsymbol{z}_{1j} = (z_{j1}, ..., z_{jn_1})^T$ and $\boldsymbol{z}_{2j} = (z_{jn_1+1}, ..., z_{jn})^T$, $j = 1, ..., d$.

Lemma 1. *It holds that as* $m \to \infty$

$$\boldsymbol{e}_{1n_1}^T \frac{\boldsymbol{W}_1^T \boldsymbol{W}_2}{(n_1 n_2)^{1/2} \lambda_{r(A)}} \boldsymbol{e}_{2n_2} = o_p(1)$$

under (2).

Proof. We write that

$$\boldsymbol{e}_{1n_1}^T \frac{\boldsymbol{W}_1^T \boldsymbol{W}_2}{(n_1 n_2)^{1/2} \lambda_{r(A)}} \boldsymbol{e}_{2n_2} = \boldsymbol{e}_{1n_1}^T \frac{\sum_{j=1}^d \lambda_{j(W)} \boldsymbol{z}_{1j} \boldsymbol{z}_{2j}^T}{(n_1 n_2)^{1/2} \lambda_{r(A)}} \boldsymbol{e}_{2n_2}.$$

Then, by using Lemma 4 given in Yata and Aoshima [9], we can conclude the result. □

Lemma 2. *It holds that as* $m \to \infty$

$$\boldsymbol{e}_{1n_1}^T \frac{\boldsymbol{A}_1^T \boldsymbol{W}_2}{n_2^{1/2} \lambda_{r(A)}} \boldsymbol{e}_{2n_2} = o_p(1) \quad \textit{and} \quad \boldsymbol{e}_{1n_1}^T \frac{\boldsymbol{W}_1^T \boldsymbol{A}_2}{n_1^{1/2} \lambda_{r(A)}} \boldsymbol{e}_{2n_2} = o_p(1)$$

under (2) and (5).

Proof. We first consider the first result of Lemma 2. We note that

$$|\boldsymbol{e}_{1n_1}^T \boldsymbol{A}_1^T \boldsymbol{W}_2 \boldsymbol{e}_{2n_2}| \leq \sum_{i=1}^r \lambda_{i(A)}^{1/2} |\boldsymbol{u}_{i(A)}^T \boldsymbol{W}_2 \boldsymbol{e}_{2n_2}|. \tag{6}$$

We write that $\boldsymbol{u}_{i(A)}^T \boldsymbol{W}_2 \boldsymbol{e}_{2n_2} = \sum_{k=1}^{n_2} e_{2k} \boldsymbol{w}_{n_1+k}^T \boldsymbol{u}_{i(A)}$. Note that $\lambda_{1(W)} = o(\lambda_{r(A)})$ as $m \to \infty$ from (2). By using Markov's inequality, for any $\tau > 0$ and $i = 1, ..., r$, we have that as $m \to \infty$

$$P\Big(\sum_{k=1}^{n_2} (\boldsymbol{w}_{n_1+k}^T \boldsymbol{u}_{i(A)})^2 / n_2 \geq \tau \lambda_{r(A)} \Big) \leq \frac{E\{\sum_{k=1}^{n_2} (\boldsymbol{w}_{n_1+k}^T \boldsymbol{u}_{i(A)})^2\}}{\tau n_2 \lambda_{r(A)}} = \frac{\boldsymbol{u}_{i(A)}^T \boldsymbol{\Sigma}_W \boldsymbol{u}_{i(A)}}{\tau \lambda_{r(A)}}$$
$$\leq \frac{\lambda_{1(W)}}{\tau \lambda_{r(A)}} = o(1)$$

from the fact that $\boldsymbol{u}_{i(A)}^T \boldsymbol{\Sigma}_W \boldsymbol{u}_{i(A)} \leq \lambda_{1(W)}$. Then, by noting that

$$\Big| \sum_{k=1}^{n_2} e_{2k} (\boldsymbol{w}_{n_1+k}^T \boldsymbol{u}_{i(A)}) / n_2^{1/2} \Big| \leq \Big\{ \sum_{k=1}^{n_2} e_{2k}^2 \Big\}^{1/2} \Big\{ \sum_{k=1}^{n_2} (\boldsymbol{w}_{n_1+k}^T \boldsymbol{u}_{i(A)})^2 / n_2 \Big\}^{1/2}$$
$$= \Big\{ \sum_{k=1}^{n_2} (\boldsymbol{w}_{n_1+k}^T \boldsymbol{u}_{i(A)})^2 / n_2 \Big\}^{1/2} = o_p(\lambda_{r(A)}^{1/2}), \tag{7}$$

from (6), we can conclude the first result. Similarly, we can conclude the second result. The proof is completed. □

Proof of Theorem 4.1. We write that for $j = 1, ..., r$

$$\frac{\tilde{\lambda}_j}{\lambda_{j(A)}} = \tilde{\boldsymbol{v}}_{1j}^T \frac{\boldsymbol{S}_{D(1)}}{\lambda_{j(A)}} \tilde{\boldsymbol{v}}_{2j} = \tilde{\boldsymbol{v}}_{1j}^T \frac{(n^{1/2}\boldsymbol{A}_1 + \boldsymbol{W}_1)^T (n^{1/2}\boldsymbol{A}_2 + \boldsymbol{W}_2)}{(n_1 n_2)^{1/2} \lambda_{j(A)}} \tilde{\boldsymbol{v}}_{2j}. \tag{8}$$

Then, by combining (8) with Lemmas 1 and 2, under (2) and (5), it holds that for $j = 1, ..., r$

$$\begin{aligned}\frac{\tilde{\lambda}_j}{\lambda_{j(A)}} &= 2\{1 + o(1)\}\tilde{\boldsymbol{v}}_{1j}^T \frac{\boldsymbol{A}_1^T \boldsymbol{A}_2}{\lambda_{j(A)}} \tilde{\boldsymbol{v}}_{2j} + o_p(1) \\ &= 2\{1 + o(1)\}\tilde{\boldsymbol{v}}_{1j}^T \frac{\sum_{j=1}^r \lambda_{j(A)} \boldsymbol{v}_{1j(A)} \boldsymbol{v}_{2j(A)}^T}{\lambda_{j(A)}} \tilde{\boldsymbol{v}}_{2j} = 1 + o_p(1).\end{aligned}$$

as $m \to \infty$. Thus, we have that

$$2^{1/2} \tilde{\boldsymbol{v}}_{ij}^T \boldsymbol{v}_{ij(A)} = 1 + o_p(1) \quad \text{for all } i, j. \tag{9}$$

It concludes the results. □

Proof of Theorem 4.2. We write that for all i, j

$$\boldsymbol{u}_{j(A)}^T \tilde{\boldsymbol{u}}_{ij} = \frac{n^{1/2} \lambda_{j(A)}^{1/2} \boldsymbol{v}_{ij(A)}^T + \boldsymbol{u}_{j(A)}^T \boldsymbol{W}_i}{(n_i \tilde{\lambda}_j)^{1/2}} \tilde{\boldsymbol{v}}_{ij}.$$

Then, from Theorem 4.1, (7) and (9) under (2) and (5), it holds that for all i, j

$$\boldsymbol{u}_{j(A)}^T \tilde{\boldsymbol{u}}_{ij} = 1 + o_p(1)$$

as $m \to \infty$. It concludes the result. □

Acknowledgements

Research of the first author was partially supported by Grant-in-Aid for Young Scientists (B), Japan Society for the Promotion of Science (JSPS), under Contract Number 26800078. Research of the second author was partially supported by Grants-in-Aid for Scientific Research (A) and Challenging Exploratory Research, JSPS, under Contract Numbers 15H01678 and 26540010.

References

[1] M. Aoshima, K. Yata, Two-stage procedures for high-dimensional data, Sequential Anal. (*Editor's special invited paper*) 30 (2011) 356-399.

[2] M. Aoshima, K. Yata, Invited review article: Statistical inference in high-dimension, low-sample-size settings, Sugaku 65 (2013) 225-247.

[3] M. Aoshima, K. Yata, The JSS research prize lecture: Effective methodologies for high-dimensional data, J. Japan Statist. Soc. Ser. J 43 (2013) 123-150.

[4] S. Jung, J. S. Marron, PCA consistency in high dimension, low sample size context, Ann. Statist. 37 (2009) 4104-4130.

[5] W. Murayama, K. Yata, M. Aoshima, Reconstruction of a signal matrix for high-dimension, low-sample-size data, RIMS Koukyuroku 1954 (2015) 23-31.

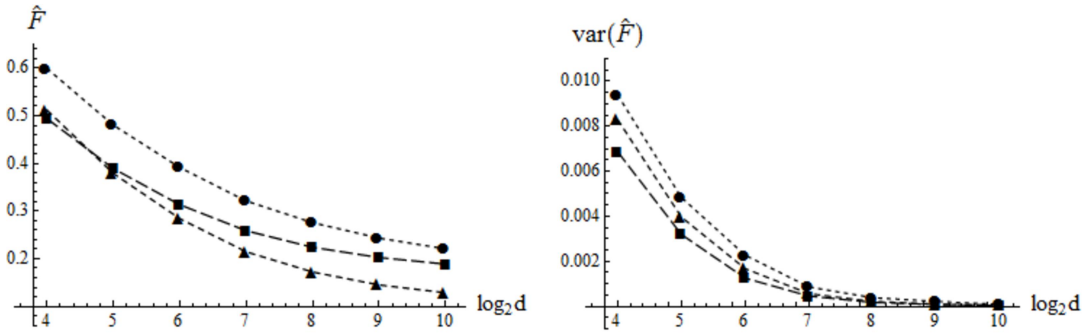
[6] A. Shabalin, A. Nobel, Reconstruction of a low-rank matrix in the presence of Gaussian noise, J. Multivariate Anal. 118 (2013) 67-76.

[7] K. Yata, M. Aoshima, Effective PCA for high-dimension, low-sample-size data with singular value decomposition of cross data matrix, J. Multivariate Anal. 101 (2010) 2060-2077.

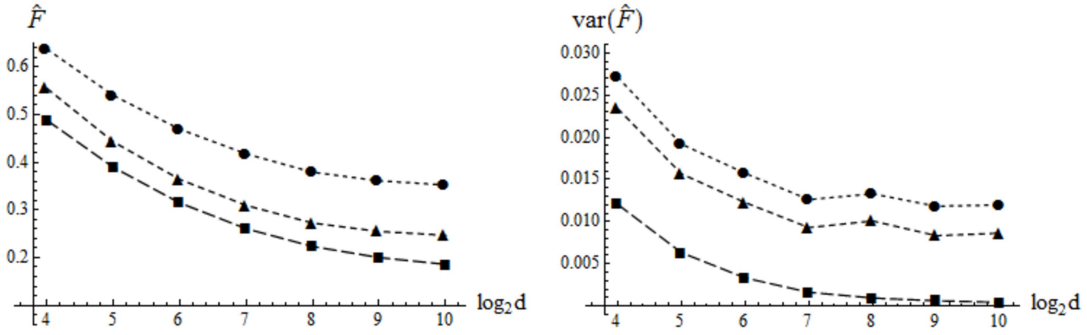
[8] K. Yata, M. Aoshima, Effective PCA for high-dimension, low-sample-size data with noise reduction via geometric representations, J. Multivariate Anal. 105 (2012) 193-215.

[9] K. Yata, M. Aoshima, PCA consistency for the power spiked model in high-dimensional settings, J. Multivariate Anal. 122 (2013) 334-354.

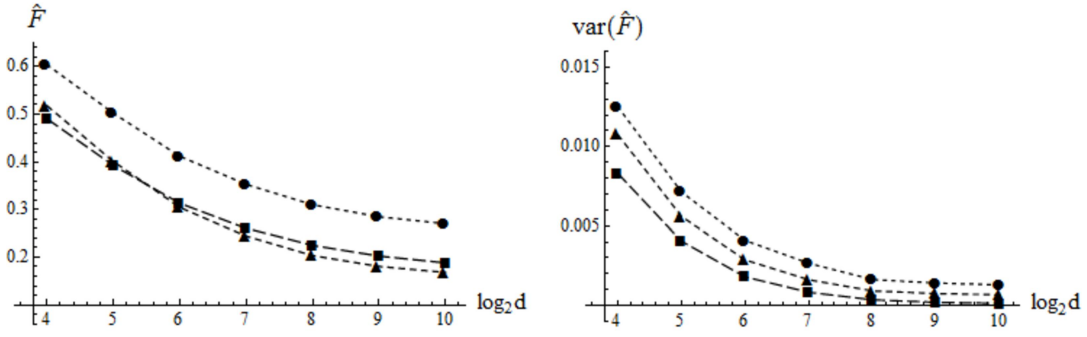
[10] K. Yata, M. Aoshima, Reconstruction of a high-dimensional low-rank matrix, Electron. J. Stat., accepted.



(a) $N_d(\mathbf{0}, \mathbf{\Sigma}_W)$



(b) $t_d(\mathbf{0}, \mathbf{\Sigma}_W, \nu)$ with $\nu = 10$



(c) $t_d(\mathbf{0}, \mathbf{\Sigma}_W, \nu)$ with $\nu = 30$

Figure 1: The behaviors of three estimates, $\hat{\boldsymbol{A}}$ denoted by ●, $\acute{\boldsymbol{A}}$ denoted by ▲ and $\tilde{\boldsymbol{A}}$ denoted by ■. The values of $\hat{F}$, $\acute{F}$ and $\tilde{F}$ are given in the left panels and their sample variances, $\mathrm{var}(\hat{F})$, $\mathrm{var}(\acute{F})$ and $\mathrm{var}(\tilde{F})$, are given in the right panels.